Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2007, Article ID 13601, 11 pages
doi:10.1155/2007/13601

Research Article

A Robust Statistical-Based Speaker's Location Detection Algorithm in a Vehicular Environment

Jwu-Sheng Hu, Chieh-Cheng Cheng, and Wei-Han Liu

Department of Electrical and Control Engineering, National Chiao Tung University, Hsinchu 300, Taiwan

Received 1 May 2006; Revised 27 July 2006; Accepted 26 August 2006

Recommended by Aki Harma

This work presents a robust speaker's location detection algorithm using a single linear microphone array that is capable of detecting multiple speech sources under the assumption that there exist nonoverlapped speech segments among sources. Namely, the overlapped speech segments are treated as uncertainty and are not used for detection. The location detection algorithm is derived from a previous work (2006), where Gaussian mixture models (GMMs) are used to model location-dependent and content and speaker-independent phase difference distributions. The proposed algorithm is proven to be robust against the complex vehicular acoustics including noise, reverberation, near-filed, far-field, line-of-sight, and non-line-of-sight conditions, and microphones' mismatch. An adaptive system architecture is developed to adjust the Gaussian mixture (GM) location model to environmental noises. To deal with unmodeled speech sources as well as overlapped speech signals, a threshold adaptation scheme is proposed in this work. Experimental results demonstrate high detection accuracy in a noisy vehicular environment.

1. INTRODUCTION

Electronic systems, such as mobile phones, global positioning systems (GPS), CD or VCD players, air conditioners, and so forth, are becoming increasingly popular in vehicles. Intelligent hands-free interfaces, including human-computer interaction (HCI) interfaces [1–3] with speech recognition, have recently been proposed due to concerns over driving safety and convenience. Speech recognition suffers from environmental noises, explaining why speech enhancement approaches using multiple microphones [4–7] have been introduced to purify speech signals in noisy environments. For example, in vehicle applications, a driver may wish to exert a particular authority in manipulating the in-car electronic systems. Additionally, for speech signal purification, a better receiving beam using a microphone array can be formed to suppress the environmental noises if the speaker's location is known.

The concept of employing a microphone array to localize sound source has been developed over 30 years [8–15]. However, most methods do not yield satisfactory results in highly reverberating, scattering or noisy environments, such as the phase correlation methods shown in [16]. Consequently, Brandstein and Silverman proposed Tukey's Biweight to the weighting function to overcome the reflection effect [17]. Additionally, histogram-based time-delay of arrival (TDOA) estimators [18–20] have been proposed for low-SNR conditions. Ward and Williamson [21] developed a particle filter beamformer to solve the reverberation problem and Potamitis et al. [22] proposed a probabilistic data association (PDA) technique to conquer these estimation errors. On the other hand, Chen et al. [23] derived the parametric maximum likelihood (ML) solution to detect speaker's location under both near-filed and far-filed conditions. To improve the computational efficiency of the ML, Chung et al. [24] proposed two recursive expectation and maximization (EM) algorithms to locate speaker. Moreover, microphones' mismatch problem is another issue for speaker's location detection [25, 26]. If the microphones are not mutually matched, then the phase difference information among microphones may be distorted. However, prematched microphones are relatively expensive and mismatched microphones are difficult to calibrate accurately since the characteristics of microphones change with the sound directions. Except for the issues mentioned above, a location detection method that can deal with the non-line-of-sight condition, which is common in vehicular environments, is necessary.

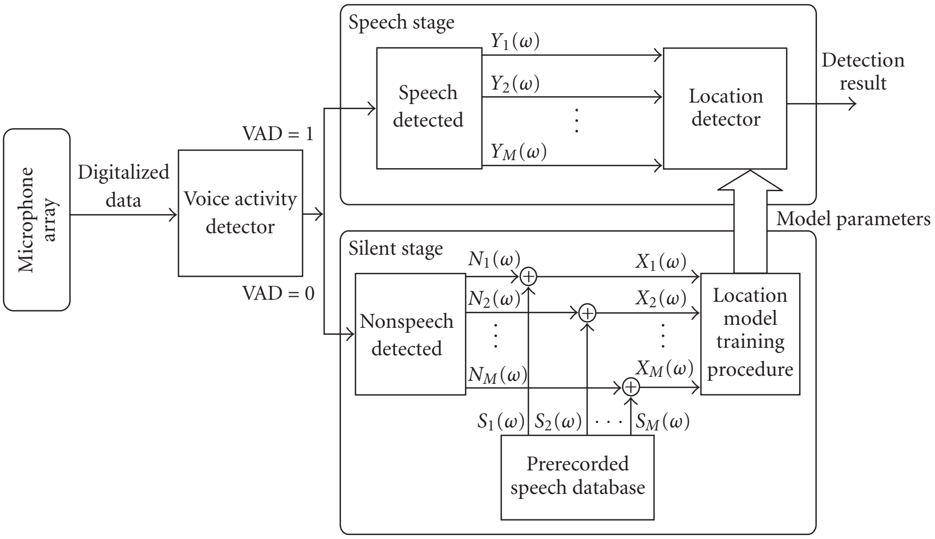


Figure 1: Overall system architecture.

Our previous work [27] utilizes Gaussian mixture model (GMM) [28] to model the phase difference distributions of the desired locations as location-dependent features for speaker's location detection. The proposed method in [27] is able to overcome the nonideal properties mentioned above and the experimental results indicate that the GMM is very suitable for modeling these distributions under both non-line-of-sight and line-of-sight conditions. Additionally, the proposed system architecture can adapt the Gaussian mixture (GM) location models to the changes in online environmental noises even under low-SNR conditions. Although the work in [27] proved to be practical in vehicular environments, it still has several issues to be solved.

First, the work in [27] assumed that the speech signal is emitted from one of the previously modeled locations. In practice, we may not want to or could not model all positions. In this case, an unexpected speech signal which is not emitted from one of the modeled locations, such as the radio broadcasting from the in-car audio system and the speaker's voices from unmodeled locations, could trigger the voice activity detector (VAD) in the system architecture, resulting in an incorrect detection of the speaker location. Second, if the speech signals from various modeled locations are mixed together (i.e., the speech signals are overlapped speech segments), then the received phase difference distribution becomes an unmodeled distribution, leading to a detection error. Therefore, this work proposes a threshold-based location detection approach that utilizes the training signals and the trained GM location model parameters to determine a suitable length of testing sequence and then obtain a threshold of the *a posteriori* probability for each location to resolve the two issues. Experimental results show that the speaker's location can be accurately detected and demonstrate that sound sources from unmodeled locations and multiple modeled locations can be discovered, thus preventing the detection error.

The remainder of this work is organized as follows. Section 2 discusses the system architecture and the relationship between the selected frequency and microphone pairs. Section 3 presents the training procedure of the proposed GM location model and the location detection method. Section 4 shows the detection performance in single and multiple speakers' cases, and the cases of radio broadcasting and speech from unmodeled locations. Conclusions are made in Section 5.

2. SYSTEM ARCHITECTURE AND MICROPHONE PAIRS SELECTION

2.1. Overall system architecture

Figure 1 illustrates the overall system architecture, which is separated into two stages, namely, the silent and speech stages, by a VAD [29, 30] that identifies speech from the received signals. Before the proposed system is processed online, a set of prerecorded speech signals are required to obtain *a priori* information between speakers and the microphone array. The prerecorded speech signals in the silent stage in Figure 1 are collected when the environment is quiet and the speakers are at the desired locations. In practice, the speakers voice several sentences and move around the desired locations slightly to simulate the practical condition and obtain an effective recording. Consequently, the pre-recorded speech signals contain both the characteristics of the microphones and the acoustical characteristics of the desired locations. After collecting the pre-recorded speech signals, the system switches automatically between the silent and speech stages according to the VAD result. If the VAD result equals to zero, indicating that speakers are silent, then the system switches to the silent stage. On the other hand, the system switches to the speech stage when the VAD result equals to one.

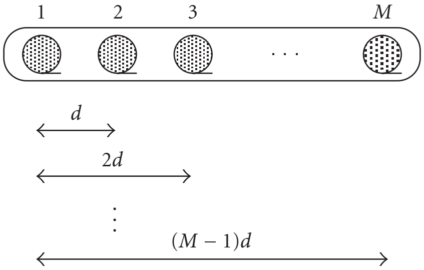


Figure 2: Uniform linear microphone array geometry.

Environmental noises without speech are recorded online in the silent stage. Given that the environmental noises are assumed to be additive, the signals received when a speaker is talking in a noisy vehicular environment can be expressed as a linear combination of the speech signal and the environmental noises. Therefore, in this stage, the system combines the online recorded environmental noise, $N_1(\omega),\dots,N_M(\omega)$, and the pre-recorded speech signals, $S_1(\omega),\dots,S_M(\omega)$, to construct the training signals, $X_1(\omega),\dots,X_M(\omega)$, where M denotes the number of microphones. The training signal is transmitted to the location model training procedure described in Section 3 to extract the corresponding phase differences and then derive the GM location models. Since the acoustical characteristics of the environmental noises may change, the GM location model parameters are updated in this stage to ensure the detection accuracy and robustness. In the speech stage, the GM location model parameters derived from the silent stage are duplicated into the location detector to detect the speaker's location.

2.2. Frequency band divisions based on a uniform linear microphone array

With the increase of the distances between microphones, the phase differences of the received signals become more significant. However, the aliasing problem occurs when this distance exceeds half of the minimum wavelength of the received signal [31]. Therefore, the distance between pairs of microphones is chosen according to the selected frequency band to obtain representative phase differences to enhance the accuracy of location detection and prevent aliasing.

Figure 2 illustrates a uniform linear microphone array with M microphones and distance d. According to the geometry, the processed frequency range is divided into $(M-1)$ bands listed in Table 1, where m denotes the mth microphone; b represents the band number, ν denotes the sound velocity, and J_b is the number of microphone pairs in the band of b. The phase differences measured by the microphone pairs at each frequency component, ω (belonging to a specific band, b) are utilized to generate a GM location model with the dimension of J_b. An example of the frequency band selection can be found in Section 4.

3. GAUSSIAN MIXTURE LOCATION MODEL TRAINING PROCEDURE AND LOCATION DETECTION METHOD

3.1. GM location model description

If the GM location model at location l is represented by the parameter $\boldsymbol{\lambda}(l) = \{\boldsymbol{\lambda}(\omega,b,l)\}|_{b=1}^{M-1}$, then a group of L GM location models can be represented by the parameters, $\{\boldsymbol{\lambda}(1),\dots,\boldsymbol{\lambda}(L)\}$. A Gaussian mixture density in the band b at location l can be denoted as a weighted sum of N Gaussian component densities:

$$G_b(\boldsymbol{\theta}_X(\omega,b,l) \mid \boldsymbol{\lambda}(\omega,b,l)) = \sum_{i=1}^{N} \rho_i(\omega,b,l) g_i(\boldsymbol{\theta}_X(\omega,b,l)), \tag{1}$$

where $\rho_i(\omega,b,l)$ is the ith mixture weight, $g_i(\boldsymbol{\theta}_X(\omega,b,l))$ denotes the ith Gaussian component density, and $\boldsymbol{\theta}_X(\omega,b,l) = [\theta_X(\omega,1,l)\cdots\theta_X(\omega,J_b,l)]^T$ is a J_b-dimensional training phase difference vector derived from the training signals, $X_1(\omega),\dots,X_M(\omega)$, as shown in the following equation:

$$\theta_X(\omega,j,l) = \text{phase}\,(X_{j+M-J_b}(\omega)) - \text{phase}\,(X_j(\omega)) \quad \text{with } 1 \le j \le J_b. \tag{2}$$

The GM location model parameter in the band b at location l, $\boldsymbol{\lambda}(\omega,b,l)$, is constructed by the mean matrix, covariance matrices, and mixture weights vector from N Gaussian component densities

$$\boldsymbol{\lambda}(\omega,b,l) = \{\boldsymbol{\rho}(\omega,b,l), \boldsymbol{\mu}(\omega,b,l), \boldsymbol{\Sigma}(\omega,b,l)\}, \tag{3}$$

where $\boldsymbol{\rho}(\omega,b,l) = [\rho_1(\omega,b,l)\cdots\rho_N(\omega,b,l)]$ denotes the mixture weights vector in the band b at location l. $\boldsymbol{\mu}(\omega,b,l) = [\boldsymbol{\mu}_1(\omega,b,l)\cdots\boldsymbol{\mu}_N(\omega,b,l)]$ denotes the mean matrix in the band b at location l. $\boldsymbol{\Sigma}(\omega,b,l) = [\boldsymbol{\Sigma}_1(\omega,b,l)\cdots\boldsymbol{\Sigma}_N(\omega,b,l)]$ denotes the covariance matrix in the band b at location l.

The ith corresponding vector and matrix of the parameters defined above are

$$\begin{gathered}\boldsymbol{\mu}_i(\omega,b,l) = [\mu_i(\omega,1,l)\cdots\mu_i(\omega,J_b,l)]^T, \\ \boldsymbol{\Sigma}_i(\omega,b,l) = \begin{bmatrix} \sigma_i^2(\omega,1,l) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_i^2(\omega,J_b,l) \end{bmatrix}.\end{gathered} \tag{4}$$

Notably, the mixture weight must satisfy the constraint that

$$\sum_{i=1}^{N} \rho_i(\omega,b,l) = 1. \tag{5}$$

The covariance matrix, $\boldsymbol{\Sigma}_i(\omega,b,l)$, is selected as a diagonal matrix. Although the phase differences of the microphone pairs may not be statistically independent of each other, GMMs with diagonal covariance matrices have been observed to be capable of modeling the correlations within the data by increasing mixture number [32].

TABLE 1: Relationship of frequency bands to the microphone pairs.

Frequency band	Microphone pairs	The number of microphone pairs	The range of frequency band
Band 1 ($b = 1$)	$(m, m + M - 1)$ with $m = 1$	$J_b = J_1 = 1$	$0 < \omega \leq \dfrac{\nu}{2(M-1)d}$
Band 2 ($b = 2$)	$(m, m + M - 2)$ with $1 \leq m \leq 2$	$J_b = J_2 = 2$	$\dfrac{\nu}{2(M-1)d} < \omega \leq \dfrac{\nu}{2(M-2)d}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Band $M - 1$ ($b = M - 1$)	$(m, m + 1)$ with $1 \leq m \leq M - 1$	$J_b = J_{M-1} = M - 1$	$\dfrac{\nu}{4d} < \omega \leq \dfrac{\nu}{2d}$

3.2. GM location models training procedure and parameters estimation

Several techniques are available for determining the parameters of the GMM, $\{\boldsymbol{\lambda}(1), \ldots, \boldsymbol{\lambda}(L)\}$, from the received phase differences. The most popular method is the EM algorithm [33] that estimates the parameters by using an iterative scheme to maximize the log-likelihood function shown as follows:

$$\log_{10} p(\boldsymbol{\theta}_X(\omega, b, l) \mid \boldsymbol{\lambda}(\omega, b, l)) = \sum_{t=1}^{T} \log_{10} p({\boldsymbol{\theta}_X}^{(t)}(\omega, b, l) \mid \boldsymbol{\lambda}(\omega, b, l)), \tag{6}$$

where $\boldsymbol{\theta}_X(\omega, b, l) = \{{\boldsymbol{\theta}_X}^{(1)}(\omega, b, l), \ldots, {\boldsymbol{\theta}_X}^{(T)}(\omega, b, l)\}$ is a sequence of T input phase difference vectors.

The EM algorithm can guarantee a monotonic increase in the model's log-likelihood value and its iterative equations corresponding to frequency band selection can be arranged as follows.

Expectation step

$$G_b(i \mid {\boldsymbol{\theta}_X}^{(t)}(\omega, b, l), \boldsymbol{\lambda}(\omega, b, l)) = \frac{\rho_i(\omega, b, l) g_i({\boldsymbol{\theta}_X}^{(t)}(\omega, b, l))}{\sum_{i=1}^{N} \rho_i(\omega, b, l) g_i({\boldsymbol{\theta}_X}^{(t)}(\omega, b, l))}, \tag{7}$$

where $G_b(i \mid {\boldsymbol{\theta}_X}^{(t)}(\omega, b, l), \boldsymbol{\lambda}(\omega, b, l))$ is a *posteriori* probability.

Maximization step

(i) Estimate the mixture weights

$$\rho_i(\omega, b, l) = \frac{1}{T} \sum_{t=1}^{T} G_b(i \mid {\boldsymbol{\theta}_X}^{(t)}(\omega, b, l), \boldsymbol{\lambda}(\omega, b, l)). \tag{8}$$

(ii) Estimate the mean vector

$$\boldsymbol{\mu}_i(\omega, b, l) = \frac{\sum_{t=1}^{T} G_b(i \mid {\boldsymbol{\theta}_X}^{(t)}(\omega, b, l), \boldsymbol{\lambda}(\omega, b, l)) {\boldsymbol{\theta}_X}^{(t)}(\omega, b, l)}{\sum_{t=1}^{T} G_b(i \mid {\boldsymbol{\theta}_X}^{(t)}(\omega, b, l), \boldsymbol{\lambda}(\omega, b, l))}. \tag{9}$$

(iii) Estimate the variances

$$\sigma_i^2(\omega, j, l) = \frac{\sum_{t=1}^{T} G_b(i \mid {\boldsymbol{\theta}_X}^{(t)}(\omega, b, l), \boldsymbol{\lambda}(\omega, b, l)) {\theta_X}^{(t)^2}(\omega, j, l)}{\sum_{t=1}^{T} G_b(i \mid {\boldsymbol{\theta}_X}^{(t)}(\omega, b, l), \boldsymbol{\lambda}(\omega, b, l))} - {\mu_i}^2(\omega, j, l) \quad \text{with } 1 \leq j \leq J_b, \tag{10}$$

where $i = \{1, \ldots, N\}$.

According to the work in [27], the location can be determined by finding the GM location model which has the maximum *posteriori* probability for a given phase difference testing sequences:

$$\hat{l} = \arg\max_{1 \leq l \leq L} \sum_{b=1}^{M-1} \log_{10}[G_b(\boldsymbol{\lambda}(\omega, b, l) \mid \boldsymbol{\theta}_Y(\omega, b))] = \arg\max_{1 \leq l \leq L} \sum_{b=1}^{M-1} \log_{10} \frac{G_b(\boldsymbol{\theta}_Y(\omega, b) \mid \boldsymbol{\lambda}(\omega, b, l)) p(\boldsymbol{\lambda}(\omega, b, l))}{p(\boldsymbol{\theta}_Y(\omega, b))}, \tag{11}$$

where $\boldsymbol{\theta}_Y(\omega, b) = \{{\boldsymbol{\theta}_Y}^{(1)}(\omega, b), \ldots, {\boldsymbol{\theta}_Y}^{(Q)}(\omega, b)\}$ is a phase difference testing sequence derived from $Y_1(\omega), \ldots, Y_M(\omega)$, and Q denotes the length of the testing sequence. However, (11) only suits for the speech signals that are emitted from one of the previously modeled locations. An unexpected speech signal which is not emitted from one of the modeled locations or a speech signal combined by the signals from various modeled locations could trigger the VAD, resulting in an incorrect detection of the speaker location. Furthermore, how to find a suitable length of the testing sequence is also an important issue.

Since conversational speech contains many short pauses, Potamitis et al. [22] locates multiple speakers by detecting the direction of individual speaker when the segments are from one single speaker and other speakers are silent (i.e., nonoverlapped speech segments). Based on this concept, this work proposes a threshold in (12) to determine whether the segment originates from a modeled location, from an unmodeled location, or from simultaneously active speakers. Because each location has specific acoustical characteristics, the threshold at each location can be used to determine whether it represents the radio broadcasting or speech signals coming from unmodeled or modeled locations. This threshold identifies the segments in which probably only one speaker in a modeled location is talking, and returns a valid location detection result.

The lengths of testing sequences and thresholds can be derived using the estimated parameters of the L GM location models. The most suitable length of testing sequences at location l is denoted as $\hat{Q}(l)$, the threshold at location l is denoted as $\zeta(l)$, and the possible searching range of the length of the testing sequence is set to $[Q_-, Q_+]$. T denotes the total length of the training phase difference sequence. $\boldsymbol{\theta}_{X,Q}(\omega,b,l,t) = \{\boldsymbol{\theta}_X{}^{(t)}(\omega,b,l),\dots,\boldsymbol{\theta}_X{}^{(t+Q-1)}(\omega,b,l)\}$ is a sequence of Q training phase difference vectors, where $1 \le t \le T - Q + 1$. The threshold varies with different length of testing sequences, so $\hat{Q}(l)$ should be determined first. To obtain a representative threshold for each location, the length of testing sequence is decided first. A suitable length of testing sequence should provide a robust characteristic under the GM location model, and a clear discrimination level between the location l and the other modeled or unmodeled GM locations. Consequently, $\hat{Q}(l)$ and $\zeta(l)$ can be obtained using the following criteria:

$$\hat{Q}(l) = \arg\max_{Q_- \le Q \le Q_+} \{C(Q)\}, \tag{12}$$

where

$$\begin{aligned} C(Q) = {}& \alpha[P_-(\boldsymbol{\lambda}(l),\boldsymbol{\theta}_X(l),Q) - P_+(\boldsymbol{\lambda}(l),\boldsymbol{\theta}_X(l),Q)] \\ &+ \beta\sum_{\substack{i=1\\ i\neq l}}^{L} I[P_-(\boldsymbol{\lambda}(l),\boldsymbol{\theta}_X(l),Q) - P_+(\boldsymbol{\lambda}(i),\boldsymbol{\theta}_X(l),Q)] \\ &+ \gamma P_-(\boldsymbol{\lambda}(l),\boldsymbol{\theta}_X(l),Q) \quad \text{with } \alpha+\beta+\gamma = 1 \end{aligned} \tag{13}$$

$$\zeta(l) = P_-\frac{(\boldsymbol{\lambda}(l),\boldsymbol{\theta}_X(l),\hat{Q}(l))}{\hat{Q}(l)}, \tag{14}$$

where α, β, γ are weights and

$$I(k) = \begin{cases} k & \text{if } k \ge 0, \\ -\infty & \text{if } k < 0. \end{cases} \tag{15}$$

$P_+(\boldsymbol{\lambda}(l),\boldsymbol{\theta}_X(l),Q)$ and $P_-(\boldsymbol{\lambda}(l),\boldsymbol{\theta}_X(l),Q)$ denote the probability upper bound and lower bound when the length of the training phase difference sequence is Q. They are derived from the following equations:

$$\begin{aligned} &P_+(\boldsymbol{\lambda}(l),\boldsymbol{\theta}_X(l),Q) \\ &\quad= \max_{\forall t}\sum_{b=1}^{M-1}\log_{10}[G_b(\boldsymbol{\lambda}(\omega,b,l)) \mid \boldsymbol{\theta}_{X,Q}(\omega,b,l,t)] \\ &P_-(\boldsymbol{\lambda}(l),\boldsymbol{\theta}_X(l),Q) \\ &\quad= \min_{\forall t}\sum_{b=1}^{M-1}\log_{10}[G_b(\boldsymbol{\lambda}(\omega,b,l)) \mid \boldsymbol{\theta}_{X,Q}(\omega,b,l,t)], \end{aligned} \tag{16}$$

where

$$\begin{aligned} &\log_{10}[G_b(\boldsymbol{\lambda}(\omega,b,l) \mid \boldsymbol{\theta}_{X,Q}(\omega,b,l,t))] \\ &\quad= \log_{10}\left[\frac{G_b(\boldsymbol{\theta}_{X,Q}(\omega,b,l,t) \mid \boldsymbol{\lambda}(\omega,b,l))\,p(\boldsymbol{\lambda}(\omega,b,l))}{p(\boldsymbol{\theta}_{X,Q}(\omega,b,l,t))}\right]. \end{aligned} \tag{17}$$

The term $p(\boldsymbol{\lambda}(\omega,b,l))$ could be eliminated because $p(\boldsymbol{\lambda}(\omega,b,l))$ is independent to t and the probability $p(\boldsymbol{\theta}_{X,Q}(\omega,b,l,t))$ is the same for all t. Therefore, (16) can be rewritten as

$$\begin{aligned} &P_+(\boldsymbol{\lambda}(l),\boldsymbol{\theta}_X(l),Q) \\ &\quad= \max_{\forall t}\sum_{b=1}^{M-1}\sum_{q=0}^{Q-1}\log_{10}G_b(\boldsymbol{\theta}_X{}^{(t+q)}(\omega,b,l) \mid \boldsymbol{\lambda}(\omega,b,l)), \\ &P_-(\boldsymbol{\lambda}(l),\boldsymbol{\theta}_X(l),Q) \\ &\quad= \min_{\forall t}\sum_{b=1}^{M-1}\sum_{q=0}^{Q-1}\log_{10}G_b(\boldsymbol{\theta}_X{}^{(t+q)}(\omega,b,l) \mid \boldsymbol{\lambda}(\omega,b,l)). \end{aligned} \tag{18}$$

The first term of (13) represents the negative maximum probability variation of the trained model when the length of the training phase difference sequence is Q. As the value of this term increases, the corresponding selection of Q yields a more robust result under the trained GM location model. The second term of (13) is the sum of the probability differences of the location l versus other locations and a larger value means the corresponding selection of Q has a higher discrimination level between the location l and the other trained GM locations. Finally, a high discrimination level between the location l and other unmodeled locations can be achieved if the third term of (13) is large. Figure 3 shows the GM location model training procedure with the total location number L.

3.3. Location detection method

The location is detected as

$$\begin{aligned} \hat{l} &= \arg\max_{1\le l\le L}\frac{1}{\hat{Q}(l)}\sum_{b=1}^{M-1}\log_{10}[G_b(\boldsymbol{\lambda}(\omega,b,l) \mid \boldsymbol{\theta}_Y(\omega,b,l))] \\ &= \arg\max_{1\le l\le L}\sum_{b=1}^{M-1}\log_{10}\frac{G_b(\boldsymbol{\theta}_Y(\omega,b,l) \mid \boldsymbol{\lambda}(\omega,b,l))\,p(\boldsymbol{\lambda}(\omega,b,l))}{\hat{Q}(l)\,p(\boldsymbol{\theta}_Y(\omega,b,l))} \end{aligned} \tag{19}$$

if

$$\begin{aligned} &\zeta\left(\arg\max_{1\le l\le L}\frac{1}{\hat{Q}(l)}\sum_{b=1}^{M-1}\log_{10}[G_b(\boldsymbol{\lambda}(\omega,b,l) \mid \boldsymbol{\theta}_Y(\omega,b,l))]\right) \\ &\quad\le \max_{1\le l\le L}\frac{1}{\hat{Q}(l)}\sum_{b=1}^{M-1}\log_{10}[G_b(\boldsymbol{\lambda}(\omega,b,l) \mid \boldsymbol{\theta}_Y(\omega,b,l))], \end{aligned} \tag{20}$$

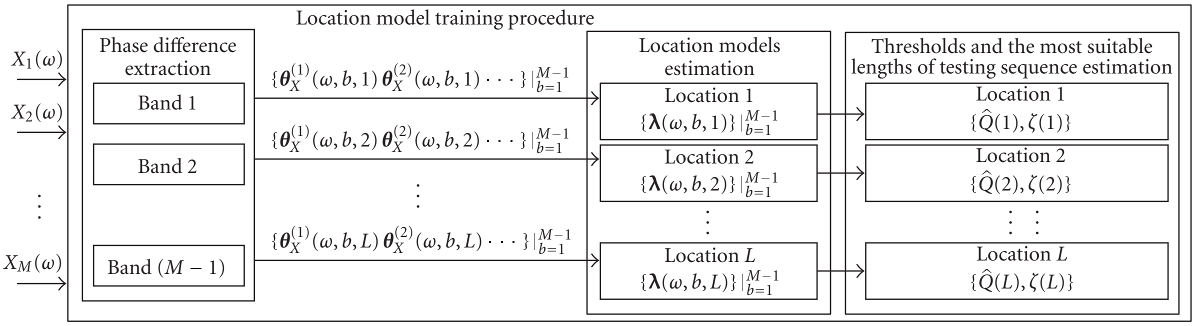

Location model training procedure
$X_1(\omega)$
$X_2(\omega)$
$X_M(\omega)$
Phase difference extraction
Band 1
Band 2
Band $(M-1)$
$\{\boldsymbol{\theta}_X^{(1)}(\omega,b,1)\,\boldsymbol{\theta}_X^{(2)}(\omega,b,1)\cdots\}|_{b=1}^{M-1}$
$\{\boldsymbol{\theta}_X^{(1)}(\omega,b,2)\,\boldsymbol{\theta}_X^{(2)}(\omega,b,2)\cdots\}|_{b=1}^{M-1}$
$\{\boldsymbol{\theta}_X^{(1)}(\omega,b,L)\,\boldsymbol{\theta}_X^{(2)}(\omega,b,L)\cdots\}|_{b=1}^{M-1}$
Location models estimation
Location 1
$\{\boldsymbol{\lambda}(\omega,b,1)\}|_{b=1}^{M-1}$
Location 2
$\{\boldsymbol{\lambda}(\omega,b,2)\}|_{b=1}^{M-1}$
Location L
$\{\boldsymbol{\lambda}(\omega,b,L)\}|_{b=1}^{M-1}$
Thresholds and the most suitable lengths of testing sequence estimation
Location 1
$\{\hat{Q}(1),\zeta(1)\}$
Location 2
$\{\hat{Q}(2),\zeta(2)\}$
Location L
$\{\hat{Q}(L),\zeta(L)\}$

Figure 3: GM location model training procedure.

where $\boldsymbol{\theta}_Y(\omega,b,l)=\{\boldsymbol{\theta}_Y{}^{(1)}(\omega,b),\dots,\boldsymbol{\theta}_Y{}^{(\hat{Q}(l))}(\omega,b)\}$ is a testing sequence derived from $Y_1(\omega),\dots,Y_M(\omega)$. If the probability densities at all locations are equally likely, then $p(\boldsymbol{\lambda}(\omega,b,l))$ could be chosen as $1/L$. The probability $p(\boldsymbol{\theta}_Y(\omega,b,l))$ is the same for all location models and then the detection rule can be rewritten as

$$\hat{l}=\arg\max_{1\le l\le L}\frac{1}{\hat{Q}(l)}\sum_{b=1}^{M-1}\sum_{q=1}^{\hat{Q}(l)}\log_{10}\left[G_b\left(\boldsymbol{\theta}_Y{}^{(q)}(\omega,b)\mid\boldsymbol{\lambda}(\omega,b,l)\right)\right] \tag{21}$$

if

$$\begin{aligned}\zeta\Bigg(&\arg\max_{1\le l\le L}\frac{1}{\hat{Q}(l)}\sum_{b=1}^{M-1}\sum_{q=1}^{\hat{Q}(l)}\log_{10}\left[G_b\left(\boldsymbol{\theta}_Y{}^{(q)}(\omega,b)\mid\boldsymbol{\lambda}(\omega,b,l)\right)\right]\Bigg)\\&\le\max_{1\le l\le L}\frac{1}{\hat{Q}(l)}\sum_{b=1}^{M-1}\sum_{q=1}^{\hat{Q}(l)}\log_{10}\left[G_b\left(\boldsymbol{\theta}_Y{}^{(q)}(\omega,b)\mid\boldsymbol{\lambda}(\omega,b,l)\right)\right].\end{aligned} \tag{22}$$

If the value of

$$\max_{1\le l\le L}\sum_{b=1}^{M-1}\sum_{q=1}^{\hat{Q}(l)}\log_{10}\frac{\left[G_b\left(\boldsymbol{\theta}_Y{}^{(q)}(\omega,b)\mid\boldsymbol{\lambda}(\omega,b,l)\right)\right]}{\hat{Q}(l)} \tag{23}$$

is not larger than the corresponding threshold, then the segments may contain speech components that come simultaneously from multiple modeled locations or from unmodeled locations.

4. EXPERIMENTAL RESULTS

The experiment was performed in a minivan with six seats [34] ($L=6$). Figure 4 shows the locations of the six in-car loudspeakers and the locations that are tested for the experiment. The first six locations correspond to modeled locations, and the radio broadcasting emits from the six in-car loudspeakers, locations no. 7, 8, and 9 correspond to unmodeled locations. A uniform linear array of six off-the-shelf, low-cost and noncalibrated microphones with 5 cm spacing is mounted in front of location no. 2. Additionally, the distance between the microphone array and the mouth of the speaker who sits in location no. 2 is about 0.62 m. In this experiment, locations no. 1 and 2 are in the near-field condition, and the signals from locations no. 3 and 5 are regarded as the far-field source according to the definition in [35]. Moreover, locations no. 4 and 6 are under the non-line-of-sight condition because the direct paths to the microphone array are sheltered by the speaker at location no. 2. The sampling rate is 8 kHz, and the A/D resolution is 16 bits. The processing window for calculating phase differences contains 256 zero-padded samples, and 32 milliseconds speech signals (512 samples in total). All windows are closed during the experiment to protect the microphones from saturation, and the cabinet temperature was set to 24°C using the in-car air conditioner.

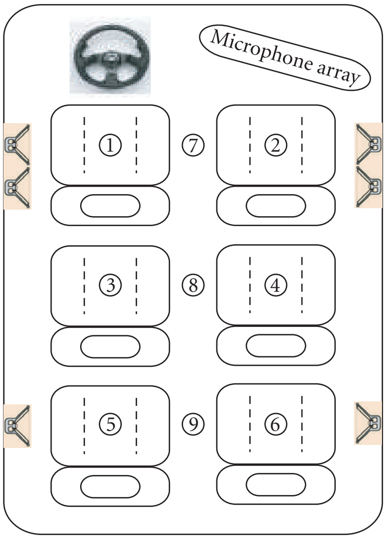
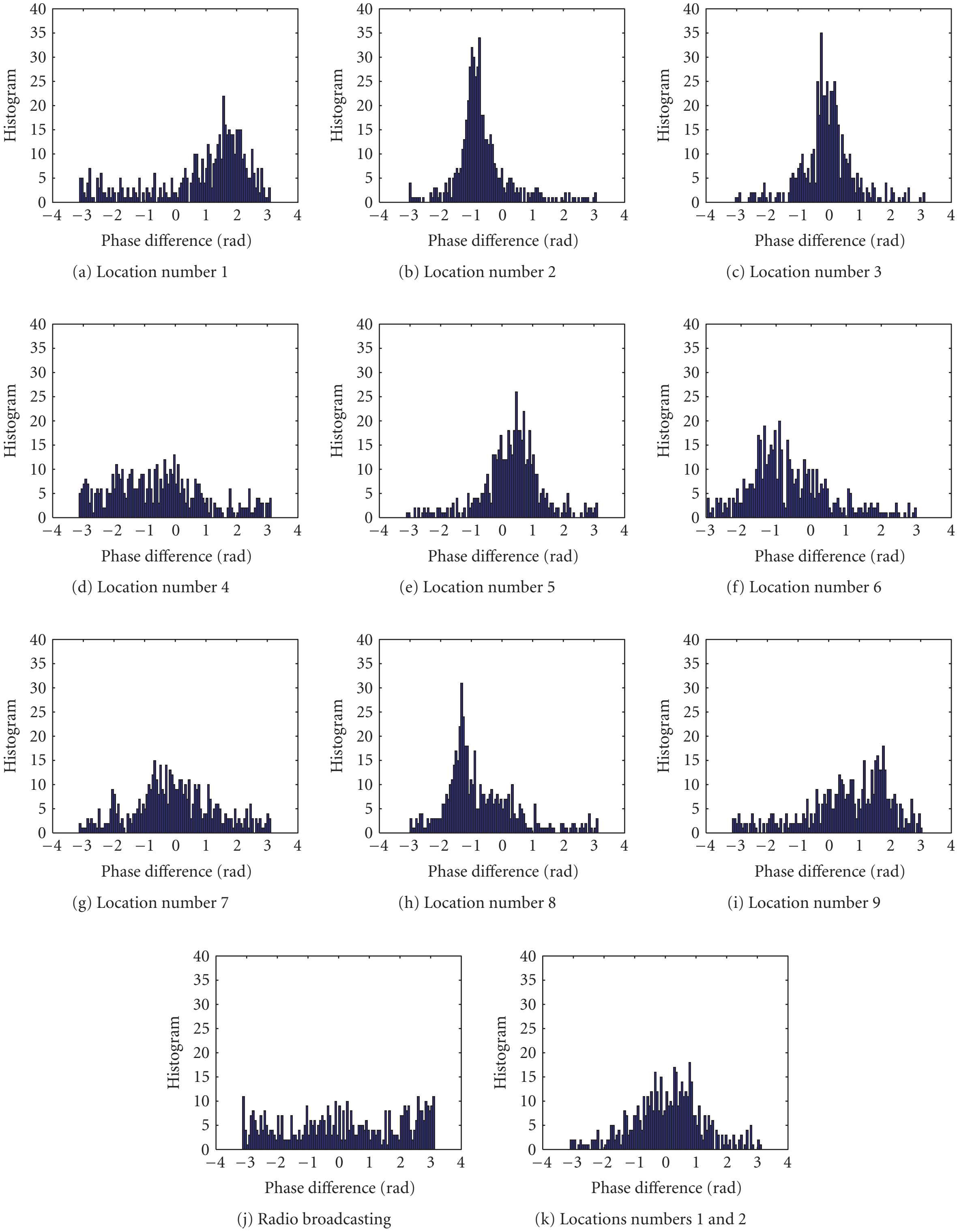


Figure 4: Locations number of the seats.

Figure 5 depicts the histograms of phase differences from individual location, and the radio broadcasting between the third and sixth microphones at the frequency of 921.875 Hz



(a) Location number 1

(b) Location number 2

(c) Location number 3

(d) Location number 4

(e) Location number 5

(f) Location number 6

(g) Location number 7

(h) Location number 8

(i) Location number 9

(j) Radio broadcasting

(k) Locations numbers 1 and 2

Figure 5: Various histograms of phase differences.

which is in the third frequency band. The histogram of phase difference in an overlapped speech segment derived when two passengers at locations no. 1 and 2 speak simultaneously is also shown in Figure 5. These phase differences are acquired when the environment is quiet. Due to the complex propagation behavior of speech signal and room acoustics, the phase difference obtained from a fixed location is a distribution instead of a fixed value. As shown in Figure 5, these

Table 2: SNR ranges at various speeds.

Speed (km/h)	SNR ranges (dB)				
	Multiple speakers at locations no. 1 to 6 (1–5 speakers)	Radio broadcasting	Single speaker at location no. 7	Single speaker at location no. 8	Single speaker at location no. 9
Speed = 0 km/h	10.81–18.15 dB	13.10 dB	14.96 dB	13.18 dB	17.31 dB
Speed = 20 km/h	5.62–12.96 dB	7.20 dB	10.15 dB	9.37 dB	11.50 dB
Speed = 40 km/h	0.19–7.54 dB	2.18 dB	4.53 dB	2.76 dB	6.89 dB
Speed = 60 km/h	−0.54–6.81 dB	1.75 dB	3.81 dB	2.03 dB	5.16 dB
Speed = 80 km/h	−5.32–2.02 dB	−3.04 dB	−0.98 dB	−2.76 dB	1.37 dB
Speed = 100 km/h	−7.28–0.07 dB	−5.99 dB	−2.93 dB	−4.71 dB	−0.58 dB

Table 3: The frequency bands correspond to the microphone pairs.

Frequency band	Microphone pairs	The number of microphone pairs	The range of frequency band
Band 1 ($b = 1$)	$(1, 6)$	$J_1 = 1$	(0 Hz, 680 Hz]
Band 2 ($b = 2$)	$(1, 5)$; $(2, 6)$	$J_2 = 2$	(680 Hz, 850 Hz]
Band 3 ($b = 3$)	$(1, 4)$; $(2, 5)$; $(3, 6)$	$J_3 = 3$	(850 Hz, 1100 Hz]
Band 4 ($b = 4$)	$(1, 3)$; $(2, 4)$; $(3, 5)$; $(4, 6)$	$J_4 = 4$	(1100 Hz, 1700 Hz]
Band 5 ($b = 5$)	$(1, 2)$; $(2, 3)$; $(3, 4)$; $(4, 5)$; $(5, 6)$	$J_5 = 5$	(1700 Hz, 3400 Hz]

phase difference distributions are quite different, as indicated by several research reports [36, 37]. Even locations no. 2, 4, and 6 which have the same angle to the microphone array cannot provide the similar distributions; given why these locations are distinguishable by pattern matching methods. Notably, the phase difference distribution from two simultaneously speaking passengers at locations no. 1 and 2 is not similar to the one from location no. 1 or 2, and thus may lead to a detection error. This phenomenon indicates that a properly selected threshold for each location can avoid the detection error caused by unmodeled locations and the overlapped speech segments.

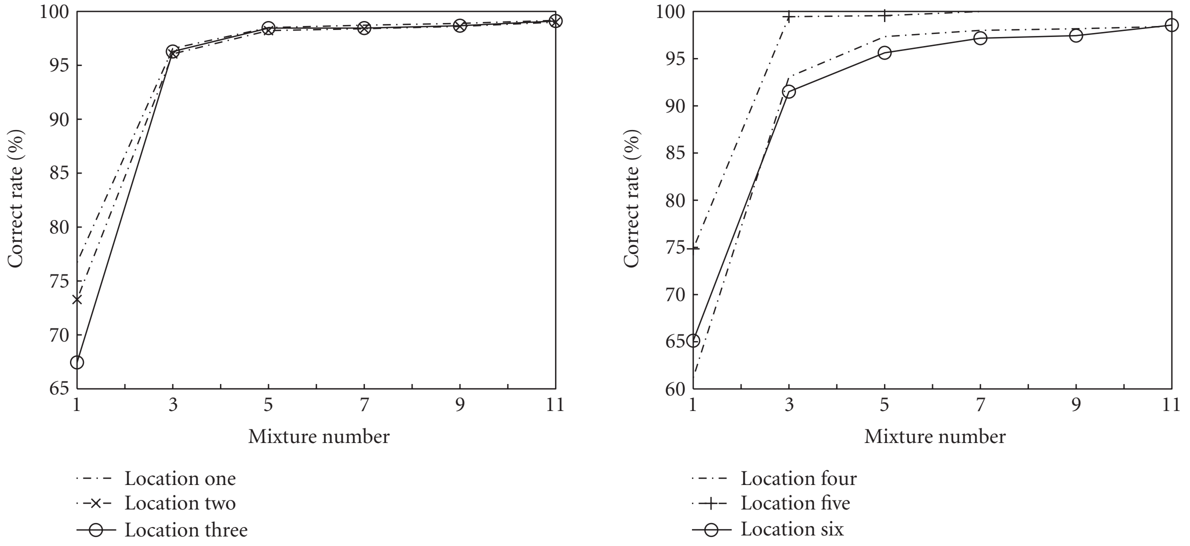
The environmental noises are varied as the vehicle runs at various speeds of 0, 20, 40, 60, 80, and 100 km/h. Table 2 lists the SNR ranges at various speeds and Table 3 presents the frequency bands that correspond to the pairs of microphones. The voice activity detection algorithm in [29] is utilized in this experiment. The total length of the training phase difference sequence T is set to 300 (3-second duration). The values of Q_-, Q_+, α, β, and γ are set to 10, 35, 0.3, 0.4, and 0.3, respectively.

The mixture number of GMM model has six choices, 1, 3, 5, 7, 9, and 11. The trial number for localization detection is 300 for each mixture number at each speed. For the condition of a single speaker, Figure 6 plots the average correct rates versus mixture numbers and indicates that a single Gaussian distribution, $M = 1$, could not yield a satisfactory performance, and that increasing the mixture number improves the performance.

Fifteen possible combinations, such as locations no. 1 and 2, locations no. 1 and 3, exist with two speakers talking. Three, four, and five speakers talking yield 20, 15, and 6 possible combinations, respectively. Table 4 lists the average error rates of these conditions with a mixture number of 11. Notably, an error is defined as a detection result that does not give the location of any of these speakers. For example, if the speech signals come from locations no. 2 and 3, then an error occurs when the detection result is neither 2 nor 3. Table 5 lists the average error rates of radio broadcasting and the speech signals coming from locations no. 7, 8, and 9 with a mixture number of 11. The error in the table is defined as the detection result pointing to one of the modeled locations. The work in [27] cannot deal with multiple speakers and unmodeled speech sources because the detection result is determined as the location with maximum a *posteriori* probability. However, the experimental results in Table 5 indicate that the method proposed in this work can successfully deal with these two conditions.

5. CONCLUSION

This work utilizes the distributions of location dependent features to construct GM location models. The proposed approach is proved to be suitable for a vehicular environment which simultaneously contains many practical issues, such as reverberation, near-filed, far-field, line-of-sight, and non-line-of-sight conditions. To prevent the detection errors caused by unmodeled location and multiple speakers' speech signal, the proposed approach computes a suitable length of testing sequence and a corresponding threshold for each modeled location. Experimental results show that the proposed approach with the suitable length of testing sequences and thresholds performs well on detecting speaker's location and on reducing the average error rates at various SNRs.



(a) Locations numbers 1 to 3

(b) Locations numbers 4 to 6

Figure 6: Average correct rates versus the mixture numbers.

Table 4: Average error rates at various speeds under multiple speakers' conditions.

Speaker number	Average error rates (%)					
	Speed = 0 km/h	Speed = 20 km/h	Speed = 40 km/h	Speed = 60 km/h	Speed = 80 km/h	Speed = 100 km/h
2	0.67%	1.11%	0.44%	0.67%	1.56%	1.78%
3	0.50%	1.00%	0.67%	0.50%	1.17%	1.83%
4	0.89%	0.89%	0.66%	0.44%	1.11%	1.56%
5	0.11%	0.05%	0%	0%	0.05%	0.11%

Table 5: Average error rates of unmodeled locations at various speeds.

Speed (km/h)	Average error rates (%)			
	Radio broadcasting	Single speaker at location no. 7	Single speaker at location no. 8	Single speaker at location no. 9
Speed = 0 km/h	0.22%	0%	0.06%	0.22%
Speed = 20 km/h	0.28%	0%	0.17%	0%
Speed = 40 km/h	0%	0%	0%	0%
Speed = 60 km/h	0.06%	0%	0%	0.33%
Speed = 80 km/h	0.28%	0.33%	0.33%	0.33%
Speed = 100 km/h	0.33%	0%	0.39%	0.67%

ACKNOWLEDGMENTS

This work is supported in part by the National Science Council of Taiwan under Grant no. NSC 93-2218-E-009-031 and the Ministry of Education, Taiwan, under Grant no. 91-1-FA06-4-4.

REFERENCES

[1] J. G. Ryan and R. A. Goubran, "Application of near-field optimum microphone arrays to hands-free mobile telephony," *IEEE Transactions on Vehicular Technology*, vol. 52, no. 2, pp. 390–400, 2003.

[2] K. Pulasinghe, K. Watanabe, K. Izumi, and K. Kiguchi, "Modular fuzzy-neuro controller driven by spoken language commands," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 34, no. 1, pp. 293–302, 2004.

[3] W. Herbordt, T. Horiuchi, M. Fujimoto, T. Jitsuhiro, and S. Nakamura, "Noise-robust hands-free speech recognition on PDAs using microphone array technology," in *Autumn Meeting of the Acoustical Society of Japan*, pp. 51–54, Sendai, Japan, September 2005.

[4] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2001.

[5] P. Aarabi and G. Shi, "Phase-based dual-microphone robust speech enhancement," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 34, no. 4, pp. 1763–1773, 2004.

[6] J.-S. Hu and C.-C. Cheng, "Frequency domain microphone array calibration and beamforming for automatic speech recognition," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E88-A, no. 9, pp. 2401–2411, 2005.

[7] S. Ahn and H. Ko, "Background noise reduction via dual-channel scheme for speech recognition in vehicular environment," *IEEE Transactions on Consumer Electronics*, vol. 51, no. 1, pp. 22–27, 2005.

[8] G. C. Carter, A. H. Nuttall, and P. G. Cable, "The smoothed coherence transform," *Proceedings of the IEEE*, vol. 61, no. 10, pp. 1497–1498, 1973.

[9] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, pp. 320–327, 1976.

[10] G. Bienvenu, "Eigensystem properties of the sampled space correlation matrix," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '83)*, vol. 8, pp. 332–335, Boston, Mass, USA, 1983.

[11] M. Wax, T.-J. Shan, and T. Kailath, "Spatio-temporal spectral analysis by eigenstructure methods," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 4, pp. 817–827, 1984.

[12] H. Wang and M. Kaveh, "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 4, pp. 823–831, 1985.

[13] J. O. Smith and J. S. Abel, "Closed-form least-squares source location estimation from range-difference measurements," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 12, pp. 1661–1669, 1987.

[14] J.-S. Hu, C.-C. Cheng, W.-H. Liu, and T. M. Su, "A speaker tracking system with distance estimation using microphone array," in *Proceedings of the IEEE/ASME International Conference on Advanced Manufacturing Technologies and Education*, pp. 485–494, Chiayi, Taiwan, August 2002.

[15] J.-S. Hu, T. M. Su, C.-C. Cheng, W.-H. Liu, and T. I. Wu, "A self-calibrated speaker tracking system using both audio and video data," in *Proceedings of the IEEE Conference on Control Applications*, vol. 2, pp. 731–735, Glasgow, Scotland, September 2002.

[16] M. Omologo and P. Svaizer, "Acoustic source location in noisy and reverberant environment using CSP analysis," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '96)*, pp. 901–904, Atlanta, Ga, USA, May 1996.

[17] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '97)*, vol. 1, pp. 375–378, Munich, Germany, April 1997.

[18] N. Strobel and R. Rabenstein, "Classification of time delay estimates for robust speaker localization," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '99)*, vol. 6, pp. 3081–3084, Phoenix, Ariz, USA, March 1999.

[19] S. Mavandadi and P. Aarabi, "Multichannel nonlinear phase analysis for time-frequency data fusion," in *Multisensor, Multisource Information Fusion: Architectures, Algorithms, and Applications 2003*, vol. 5099 of *Proceedings of SPIE*, pp. 222–231, Orlando, Fla, USA, April 2003.

[20] P. Aarabi and S. Mavandadi, "Robust sound localization using conditional time-frequency histograms," *Information Fusion*, vol. 4, no. 2, pp. 111–122, 2003.

[21] D. B. Ward and R. C. Williamson, "Particle filter beamforming for acoustic source localization in a reverberant environment," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '02)*, vol. 2, pp. 1777–1780, Orlando, Fla, USA, May 2002.

[22] I. Potamitis, H. Chen, and G. Tremoulis, "Tracking of multiple moving speakers with multiple microphone arrays," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 520–529, 2004.

[23] J. C. Chen, K. Yao, and R. E. Hudson, "Acoustic source localization and beamforming: theory and practice," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 4, pp. 359–370, 2003.

[24] P.-J. Chung, J. F. Böhme, and A. O. Hero, "Tracking of multiple moving sources using recursive EM algorithm," *EURASIP Journal on Applied Signal Processing*, vol. 2005, no. 1, pp. 50–60, 2005.

[25] B. C. Ng and C. M. S. See, "Sensor-array calibration using a maximum-likelihood approach," *IEEE Transactions on Antennas and Propagation*, vol. 44, no. 6, pp. 827–835, 1996.

[26] D. B. Ward, E. A. Lehmann, and R. C. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 826–836, 2003.

[27] J.-S. Hu, C.-C. Cheng, and W.-H. Liu, "Robust speaker's location detection in a vehicle environment using GMM models," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 36, no. 2, pp. 403–412, 2006.

[28] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[29] J. Ramírez, J. C. Segura, C. Benítez, A. De la Torre, and Á. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, no. 3-4, pp. 271–287, 2004.

[30] I. Potamitis, "Estimation of speech presence probability in the field of microphone array," *IEEE Signal Processing Letters*, vol. 11, no. 12, pp. 956–959, 2004.

[31] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, chapter 2, Springer, New York, NY, USA, 2001.

[32] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[33] G. Xuan, W. Zhang, and P. Chai, "EM algorithms of Gaussian mixture model and hidden Markov model," in *Proceedings of the IEEE International Conference on Image Processing (ICIP '01)*, vol. 1, pp. 145–148, Thessaloniki, Greece, October 2001.

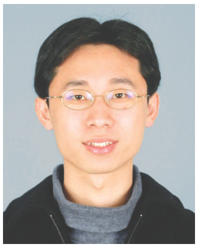
[34] Mitsubishi Motors - Savrin (http://www.sym-motor.com.tw/savrin-1.htm).

[35] J. G. Ryan and R. A. Goubran, "Near-field beamforming for microphone arrays," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '97)*, vol. 1, pp. 363–366, Munich, Germany, April 1997.

[36] D. D. Vries, E. M. Hulsebos, and J. Bann, "Spatial fluctuations in measures for spaciousness," *Journal of the Acoustical Society of America*, vol. 110, no. 2, pp. 947–954, 2001.

[37] X. Pelorson, J.-P. Vian, and J.-D. Polack, "On the variability of room acoustical parameters: reproducibility and statistical validity," *Applied Acoustics*, vol. 37, no. 3, pp. 175–198, 1992.

Jwu-Sheng Hu was born in Taipei, Taiwan, in 1962. He received the B.S. degree from the Department of Mechanical Engineering, National Taiwan University, Taiwan, in 1984, and the M.S. and Ph.D. degrees from the Department of Mechanical Engineering, University of California at Berkeley, in 1988 and 1990, respectively. He is currently a Professor in the Department of Electrical and Control Engineering, National Chiao Tung University, Taiwan, His current research interests include microphone array signal processing, active noise control, embedded system design, and robotics.



Chieh-Cheng Cheng was born in 1978. He received the B.S. and Ph.D. degrees in electrical and control engineering from National Chiao Tung University, Taiwan, in 2000 and 2006, respectively. He is the Championship of TI DSP Solutions Design Challenge in 2000 and of the national competition held by Ministry of Education Advisor Office in 2001. His research interests include sound source localization, microphone array signal processing, adaptive signal processing, pattern recognition, speech signal processing, and echo and noise cancellations.

Wei-Han Liu was born in Kaohsiung, Taiwan, in 1977. He received the B.S. and M.S. degrees in electrical and control engineering from National Chiao Tung University, Taiwan, in 2000 and 2002, respectively. He is currently a Ph.D. candidate in Department of Electrical and Control Engineering at National Chiao Tung University, Taiwan. He is the Championship of TI DSP Solutions Design Challenge in 2000 and of the national competition held by Ministry of Education Advisor Office in 2001. He is the winner of the Best Paper Award at IEEE/ASME 2002. His research interests include sound source localization, microphone array signal processing, adaptive signal processing, speech signal processing, and robot localization.